NLDR methods for high dimensional NIRS dataset: application to vineyard soils characterization

Clément Delion[1], Ludovic Journaux[1,2], Aurore Payen[1], Lucile Sautot[1,3]
Emmanuel Chevigny[4], Pierre Curmi[1,5]

1- AgroSup Dijon
26 Boulevard du Docteur Petitjean 21000 Dijon - France

2- Université de Bourgogne - UMR6306 LE2I
Avenue Alain Savary 21000 Dijon - France

3- Université de Bourgogne - UMR6282 Biogéosciences
6 Boulevard Gabriel 21000 Dijon - France

4- Université de Bourgogne - UMR6298 ArteHis
6 Boulevard Gabriel 21000 Dijon - France

5- INRA Dijon - UMR1347 Agroécologie
17 rue Sully 21000 Dijon - France

Abstract. In the context of vineyard soils characterizationn this paper explores and compare different recent Non Linear Dimensionality Reduction (NLDR) methods on a high-dimensional Near InfraRed Spectroscopy (NIRS) dataset. NLDR methods are based on k-neighborhood criterion and Euclidean and fractional distances metrics are tested. Results show that Multiscale Jensen-Shannon Embedding (Ms JSE) coupled with euclidean distance outperform all over methods. Application on data is made at global scale and at different scale of depth of soil.

1 Introduction

NIRS is an efficient spectroscopic method for the quantification and characterization of soil components. Generally NIRS analyses are focusing on some regions of the spectrum as in [8] while other studies are focussing on the all spectral signature treated by linear Principal Component Analysis (PCA) in order to identify biogenic structures [11]. Unfortunately, these high dimensional (HD) datasets are difficult to handle, as the information is often redundant and highly correlated with one another. Moreover, this HD dataset can suffer from the curse of dimensionality [3], like norm concentration and hubness. Thus, to improve the characterization performance and/or HD data visualization, it is well interesting to use NLDR techniques to transform HD data into a meaningful representation of reduced dimensionality. Some studies have tried to compare NLDR methods to linear methods, often for synthetic data such as the swissroll, but less for HD natural data. The aims of this paper is to find the best NLDR method applied to the dataset in order to characterize vineyard soils. A double variability is analyzed: an interspecific due to different sites,and an intraspecific due to the samples. Thus, we will make a global analysis, then an analysis depending on soil

depth. This paper is organized as follows.Section 2 presents the NIRS dataset sampling used on this context, the fractional metrics and an NLDR methods overview with a comparison based on a quality assessment. Section 3 presents and discusses experimental results. Section 4 draws the conclusions.

2 Materials and methods

2.1 Presentation of the NIRS dataset

NIRS dataset comes from four representative vineyard places of Burgundy: Aloxe Corton, Couchey, Maranges and Monthelie. Samples are extracted on surface and at different soil depths. For each sample (dried, screened and crushed), three NIRS acquisition are made with a FieldSpec 3 (ASD Inc.). Each spectrum scans wavelengths between 350 and 2500 nm. Therefore the dataset present a dimensionality of 2151. For each places numerous samples were taken. The gathered results are characterized by one of the four sites and the soil layer that is associated with them. Finally, we obtain 13 drillings in Aloxe Corton, 14 in Couchey, 11 in Maranges and 8 in Monthelie.

2.2 Overview of different methods of dimensional reduction

We select nine of DR methods based on their scale analysis, reflecting the compromise based by NLDR methods between global structure and preservation of neighborhood at a local scale, and their distance or similarity. We retained 7 NLDR methods completed with to 2 linear methods: the *Classical Multidimensional Scaling* (CMDS) and the *Non-metric Multidimensional Scaling* (NMDS) [6]. *Non Linear Mapping* (NLM): Sammon's mapping [6] tries to preserve the neighborhood topology of data by minimizing differences in distances between the HD space and the low-dimensiona (LD) space by the Sammon's space function. *Curvilinear Component Analysis* (CCA) [6] tries to preserve pairwise distances, but gives priority to small distances by incorporating the divergence of Bregman in its stress function [9]. *Stochastic Neighbor Embedding* (SNE) [2] is a non-linear reduction method based on similarities between points, which converts pairwise distances into probabilities that represent similarities, where the most similar points have a higher probability, and then recalculates these probabilities in the LD space and minimizes the Kullback-Leibler (KL) divergence between two distributions. *t-distributed Stochastic Neighbor Embedding* (t-SNE)[10] is similar to SNE, difference is in the calculation of the probability distributions. The SNE uses a Gaussian distribution, while the t-SNE is based on a Student distribution. *Neighbor Retrieval Visualizer* (NeRV)[12] is similar to the t-SNE. The main difference is the minimization of two dual KL divergences, that are related to precision and recall, instead of a single KL divergence. The optimization of two functions, and not only one, allows a better optimization of the divergence. *Jensen-Shannon Embedding* (JSE)[5] is based on the preservation of the neighborhood. Unlike previous methods JSE uses the Jensen-Shannon divergence instead of the KL divergence to measure the similarities between two

probability distributions. *Multiscale Jensen-Shannon Embedding* (Ms. JSE)[4] is an improvement of JSE which overcomes the problematic of the neighborhood size by taking into account multiple sample sizes, thanks to a log scale.

2.3 Fractional distance transformation

Nearest neighbor research often rely on the use of the euclidean distance. Unfortunately when data represent a high dimensional features, the euclidean distances seem to concentrate and all distances between pairs of data elements seem to be very similar. Therefore, the pertinence of the euclidean distance has been questioned in different works, and fractional distance has been proposed in order to overcome the problem of concentration phenomenon or curse of dimensionality such as in [1]. In order to test if fractional distance can improve NDLR result, we test it on our NIRS dataset.

2.4 Quality criterion used for an objective comparison

In order to compare different methods between them. We use the quality assessment criterion [7]. An evaluation based on the performance of cost functions of NLDR methods is irrelevant, due to the variability of criterions used in cost functions (mean,variance, standard deviation...). Therefore, a k-neighborhood quality function is define and we compare it to an average random projection. Then we compute the area under the curve (AUC) of representing the score which is a scale-independent quality criterion for comparing methods:

$$AUC = \frac{\sum_{K=1}^{N-2} \left\{ 100 \frac{N-1}{N-K} \left(\left[\sum_{K=1}^{N-2} \frac{|v_i^K \cap n_i^K|}{KN} \right] - \frac{K}{N-1} \right) \right\} / K}{\sum_{K=1}^{N-2} 1/K}$$

with K: number of neighbors, N: number of points, v: vector of K nearest neighbors of point i in HD space, n: vector of K nearest neighbors of point$i$ in LD space.

3 Results and discussion

3.1 Results on raw data and fractional distance

We first compared the NLDR methods on the raw NIRS dataset with the quality criterion (figure 1). Ms. JSE seems the best method with 76.8% of improvement over a random projection. The CMDS (64.1%) gives also good results but is less effective according to the quality criterion. Even if CMDS is very successful at a global scale, Ms. JSE is better due to good representation of both global and local scales. MS JSE will be used in section 3.3 for soil characterization. The use of an other metric, like fractional distances, can confirm the choice of Ms. JSE as preferential NLDR method. Some studies introduce fractional distances as an alternative of Euclidean distances to analyze dataset. So this is a good basis to confirm the previous comparison. Ms. JSE is again the best method with an improvement of 73.9% over a random projection is obtain. Unfortunaltely, the results obtain show that the fractional distances give poorer results than Euclidean distances with all methods. The only exception is with the t-SNE which loses only 0.1%.

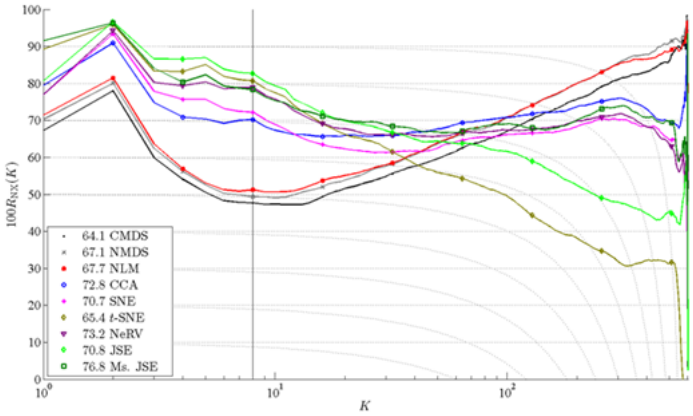

Figure 1: Comparison of DR methods with R_{NX} functions on the raw NIRS dataset

3.2 Resulting projection of the data with Ms. JSE

Following the choice of Ms. JSE as the best approach, we specifically project the NIRS data with this method (figure 2). Remember that the Aloxe Corton boreholes are numbered 1 to 13, those of Couchey from 14 to 27, those of Maranges from 28 to 38 and those of Monthelie from 39 to 46. The projection shows a data organization in clusters. We can observe that each cluster represent sampling place (it may be noted that boreholes of the same place are closed). Moreover, the analysis of the Maranges place clearly shows two clusters, representing two different types of soil on this site, which confirms the capability of Ms. JSE to discriminate different clusters and so soils characteristics.

3.3 Clustering on NIRS dataset depending on depth

In order to recognize different terroirs of Burgundy, we perfom a k-means algorithm on the same depth data points. We assumed that there were 5 clusters corresponding to terroirs. So we used the k-means method with 5 clusters as the input argument. Each figure has its own clusters at each depth. Thus, we are able to finely sign and visualize the composition of vineyard soils layers for each depth (figure 3).

3.4 Discussion

In this paper, we demonstrate the Ms. JSE efficiency. But figure 1 shows that Ms. JSE has medium performances at a global scale. In fact, Ms. JSE is the best trade-off between local and global scale: we can tolerate medium performances at global scale because local performances are very high. In some cases that involve considerating global performances, CMDS, NMDS or NLM can therefore provide best results than Ms. JSE. Thus, the superiority of Ms. JSE is proved only for our application case and cannot be considered as a general statement.

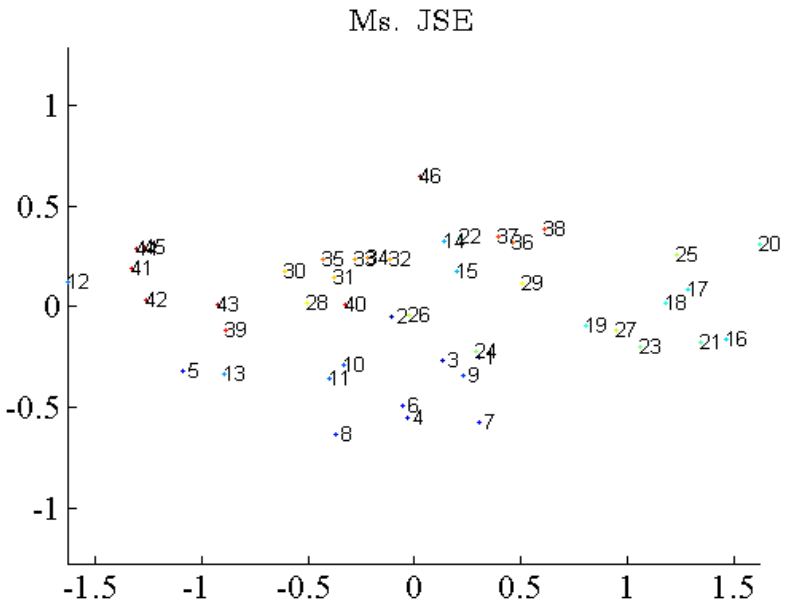

Figure 2: Results of Ms. JSE on NIRS dataset (R_{NX}: 85.3%)

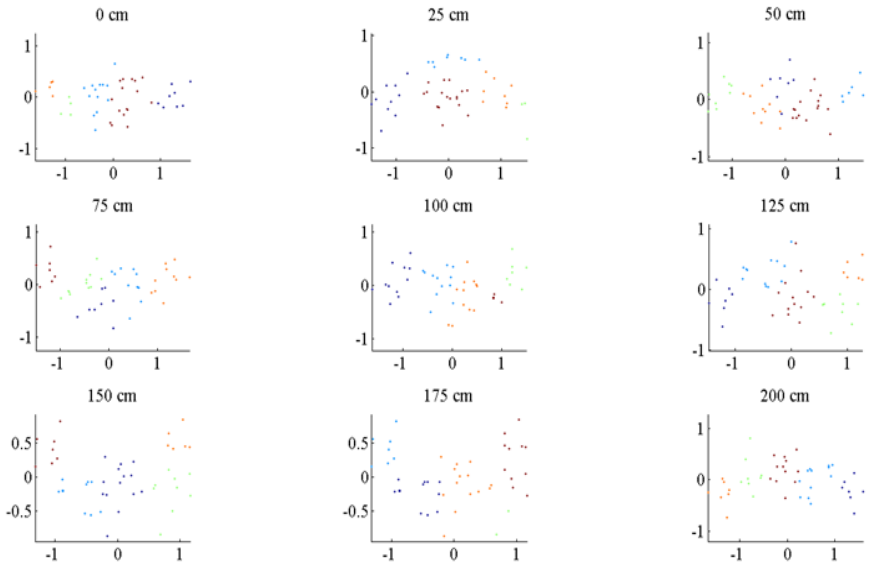

Figure 3: Results on depth data with k-means clustering (K=5)

4 Conclusion

This paper explores the comparison and application of different NLDR methods on a high-dimensional dataset in order to characterize vineyard soils by their spectral signatures. We determine that Ms. JSE with Euclidean distances is the most suitable method for our biological dataset, thanks to a quality criterion based on the k-nearest neighbor algorithm. The global analysis shows different types of vineyard soils. Finally we try a clustering in an analysis by depth and highlight 5 clusters with the k-means algorithm. Then the determination of the spectral signature of a vineyard soil permits less analysis to determine its chemical composition.

References

[1] Damien François, Vincent Wertz, Michel Verleysen, and Senior Member. The concentration of fractional distances. *IEEE Trans. on Knowledge and Data Engineering*, 19:873–886, 2007.

[2] Geoffrey E. Hinton and Sam T. Roweis. *Advances in neural information processing systems*, chapter Stochastic neighbor embedding, pages 833–840. 2002.

[3] G Hughes. On the mean accuracy of statistical pattern recognizers. *Information Theory, IEEE Transactions on*, 14(1):55–63, 1968.

[4] John A. Lee, Diego H. Peluffo-Ordonez, and Michel Verleysen. Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction. In *Proceedings of 22st European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning (ESANN)*, 2014.

[5] John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[6] John Aldo Lee and Michel Verleysen. *Nonlinear Dimensional Reduction*. Information Science and Statistics. Springer, 2007.

[7] John Aldo Lee and Michel Verleysen. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing*, 72:1431–1443, 2009.

[8] Bo Stenberg, Raphael A Viscarra Rossel, Abdul Mounem Mouazen, and Johanna Wetterlind. Chapter five-visible and near infrared spectroscopy in soil science. *Advances in agronomy*, 107:163–215, 2010.

[9] Jigang Sun, Colin Fyfe, and Malcolm Crowe. Curvilinear component analysis and bregman divergences. In *ESANN*. Citeseer, 2010.

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[11] Elena Velasquez, Céline Pelosi, Didier Brunet, Michel Grimaldi, Marlucia Martins, Ana Carolina Rendeiro, Edmundo Barrios, and Patrick Lavelle. This ped is my ped: visual separation and near infrared spectra allow determination of the origins of soil macroaggregates. *Pedobiologia*, 51(1):75–87, 2007.

[12] Jarkko Venna, Jaakko Peltonen, and Kristian Nybo et al. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.